AD-782 266

COUNTABLE STATE CONTINUOUS TIME DYNAMIC PROGRAMMING WITH STRUCTURE

Steven A. Lippman

California University

Prepared for:

Air Force Office of Scientific Research

April 1974

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR - TR - 74 - 1136	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER AD 782266
4. TITLE (and Subtitle) COUNTABLE STATE CONTINUOUS TIME DYNAMIC PROGRAMMING WITH STRUCTURE		5. TYPE OF REPORT & PERIOD COVERED Interim
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Steven A. Lippman		8. CONTRACT OR GRANT NUMBER(s) AFOSR 72-2349
9. PERFORMING ORGANIZATION NAME AND ADDRESS The University of California Western Management Science Institute Los Angeles, California 90024		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 9769-05
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM 1400 Wilson Blvd Arlington, Virginia 22209		12. REPORT DATE April 1974
		13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

A. Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

optimal control
dynamic programming
Markov Decision Processes
queueing optimization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider the problem P of maximizing the expected discounted reward earned in a continuous time Markov decision process with countable state and finite action space. (The reward rate is merely bounded by a polynomial.) By examining a sequence $\langle P_N \rangle$ of approximating prob-

#20/Abstract

lems, each of which is a semi-Markov decision process with exponential transition rate Λ_N, $\Lambda_N \nearrow \infty$, we are able to show that P is obtained as the limit of the P_N. The value in representing P as the limit of P_N is that structural properties present in each P_N persist, both in the finite and in the infinite horizon problem. Three queueing optimization models illustrating the method are given.

WESTERN MANAGEMENT SCIENCE INSTITUTE

University of California, Los Angeles

Working Paper No. 213

COUNTABLE STATE CONTINUOUS TIME DYNAMIC PROGRAMMING WITH STRUCTURE*

by

Steven A. Lippman

April, 1974

*This research was supported by the Air Force Office of Scientific Research through Grant AFOSR-72-2349a.

iii

I. INTRODUCTION

In a recent paper [7], we demonstrated the usefulness of enlarging the standard set of decision epochs in a number of semi-Markov decision processes in which the time between transitions is exponential. The approach employed in [7] was to augment the standard set of decision epochs -- usually the times the system changes state -- so that the exponential transition time has parameter Λ, independent of both the current state and the action selected. The motivation and power of this approach stem from the fact that the augmented set of decision epochs results in an n-period problem in which the expected horizon length (is n/Λ and) does not depend upon either the control policy employed or the initial state of the system. This new formulation does not typically lead to the dissipation of desirable properties -- such as monotonicity and concavity -- of the return function as is often true in the "improper" standard formulation.

Presently, we intend to make use of this technique in the context of truly continuous time problems with denumerable state space and finite action space, where by a continuous time problem we mean one in which the decision maker must select an action at each and every instant of time. Specifically, we will show that the continuous time problem $\mathcal{P}$ can be obtained as the limit of a sequence of approximating problems $\mathcal{P}_N$, where $\mathcal{P}_N$ is the obvious semi-Markov version of $\mathcal{P}$ but with exponential parameter Λ_N. Of course, we must have $\Lambda_N \nearrow \infty$. In particular, it is shown that the return in $\mathcal{P}_N$ of any measurable policy converges to its return in $\mathcal{P}$.

The advantage of obtaining P as the limit of the P_N is that certain structural properties present in each P_N persist in the limit. For example, if the optimal policy is monotone in that it uses larger actions from larger states or if the optimal return function is convex for each N, then these properties are preserved in passing to the limit. One could, of course, deterministically form a grid of decision points spaced $1/\Lambda_N$ time units apart rather than do this stochastically as we suggest. For the applications we envisage, however, use of a deterministic grid renders the approximating problems P_N themselves difficult to solve.

Previously, Miller [11] successfully treated the undiscounted finite horizon problem, and the infinite horizon problem, with and without discounting, has been covered by Kakumanu [5] and by Miller [12]. The focus and intent of this paper, however, is not to present a theory of continuous time Markov decision processes but rather to present a method or approach for dealing with problems and models whose natural formulation results in a continuous time Markov decision process; such models often possess the kind of structure necessary to render our approach applicable.

The requisite notation is introduced in section 2 while our approach and main results are presented in sections 3 and 4 for the finite and infinite horizon problems, respectively. The final section contains several applications.

II. NOTATION AND PROBLEM DEFINITION

We consider a continuous time Markov decision process with countable state space S and action space $A = \underset{s\in S}{X} A_s$ in which each coordinate A_s is finite. For convenience, we take S to be the nonnegative integers. The reward rate associated with being in state i and selecting action a is denoted by r(i,a), and the transition rate to state j from state i while employing action a is given by $q(j|i,a)$. Of course, $q(j|i,a) \geq 0$ for $j \neq i$ and $q(i|i,a) = -\Sigma_{j\neq i}\, q(j|i,a)$.

Following Miller [11] and Kakumanu [5], a policy π is simply a mapping from $S\times[0,T]$ into A, where $T \leq \infty$ is the time horizon. Thus, only deterministic memoryless rules are allowed. In addition, we require that $\pi(i,t)$, the action prescribed by π at time t from state i, be measurable in t for each i. Of particular interest are stationary and piecewise constant policies. If for each i there is an integer $n_i < \infty$ and a sequence $0 = t_0 \leq t_1 \leq \ldots \leq t_{n_i} = T$ such that $\pi(i,t)$ is constant on $[t_j,t_{j+1})$, $j = 0,1,\ldots,n_i-1$, then π is said to be <u>piecewise constant</u>. If π is piecewise constant and $n_i \equiv 1$, then π is said to be <u>stationary</u>. As we shall see, an optimal policy can be found among the class of stationary policies if $T = \infty$ and, with an additional structural condition, among the piecewise constant policies if $T < \infty$.

While our definition of a piecewise constant policy differs from that of Miller [12], they are the same if S is finite. Let F be the set of maps from S into A so that $\pi = \{f_t\}$, $f_t \in F$. Then, according to Miller, π is piecewise constant if there is a sequence $0 = t_0 < t_1 < \ldots < t_n = T < \infty$ such that $t,t' \in (t_j,t_{j+1})$ implies $f_t = f_{t'}$, $j = 0,1,\ldots,n-1$.

Miller [11] proved that for S and T finite and $\alpha = 0$ there is a piecewise constant policy that is optimal. His result does not, however, remain valid if S is countable as evidenced in the following example: Take $r(i,a) = i^2 + a$, $A_i = \{0,1\}$, and $q(i-1|i,1) = 1 = -q(i|i,1)$, $q(i|i,0) = 0$. Define $f_j \in F$ by $f_j(k) = 0$ for $k < j$ and $f_j(k) = 1$ for $k \geq j$. It is clear upon reflection that given $T < \infty$ there is an $m < \infty$ and a strictly increasing sequence $<t_j>_0^\infty$ with $t_0 = 0$ and $t_j < T$ all j such that π^* is optimal and $\pi^*(t) = f_{j+m}$ for $t \in [t_j, t_{j+1})$, $j = 0,1,\ldots$.

So that the expected return of each policy is finite, we need to make the following three assumptions. The first merely stipulates that, with probability 1, only a finite number of changes of state take place in a finite amount of time. Because our interest in such systems is motivated in large part by queueing reward systems, we do not, as is typical, require $\{r(s,a)\}$ to be bounded, but rather we impose two less restrictive assumptions (see [8]). Assumption 2 places a polynomial bound on $\max_a |r(s,a)|$ while Assumption 3 requires that movement to distant states carry small probability.

<u>Assumption 1</u>: There is a finite constant Λ such that

(1) $$\Lambda = \sup \{-q(i|i,a) : a \in A_i,\ i \in S\} .$$

<u>Assumption 2</u>: There are finite integers K and m so that for each $i \geq 0$ we have

(2) $$\max_{a \in A_i} |r(i,a)| \leq K(i \vee 1)^m .$$

Assumption 3: There is a $b < \infty$ such that for each i,

(3) $$\max_{a \epsilon A_i} -\sum_{j \neq i} (j \vee 1)^n q(j|i,a)/q(i|i,a) \leq (i + b)^n, \quad n = 1,2,\ldots,m .$$

Given a policy π, denote the total expected α-discounted reward earned on the time interval $[t,T]$ when starting at time t from state i by $V_{\alpha,t}(\pi,i)$ so that

$$V_{\alpha,t}(\pi) = \int_t^T e^{-\alpha(\xi-t)} P(t,\xi,\pi) r(\pi(\xi))d\xi ,$$

where $P(t,\xi,\pi)$ is the unique transition probability matrix associated with π. Also, define the optimal return function $V_{\alpha,t}$ by

$$V_{\alpha,t}(i) = \sup_{\pi} V_{\alpha,t}(\pi,i) .$$

A policy π^* is said to be optimal if $V_{\alpha,t}(\pi^*) = V_{\alpha,t}$, all $t \leq T$. When $T = \infty$, we simply drop the t and write V_α and $V_\alpha(\pi)$. Finally, we refer to the above as problem $\mathcal{P}$ or simply $\mathcal{P}$.

Associated with $\mathcal{P}$ is a sequence $\langle \mathcal{P}_N \rangle$ of approximating problems each of which is a semi-Markov decision process. First consider the case $T < \infty$ and set $\Lambda_N = 2^N/T$. Then by problem N we mean that 2^N-period problem with state space S, action space A, and reward function $r_{\alpha,N}$ given by $r_{\alpha,N}(s,a) = r(s,a)/(\alpha+\Lambda_N)$; the transition time is exponential with parameter Λ_N for each $(s,a) \epsilon S \times A$, and the law of motion q_N is given by

(4) $$q_N(j|i,a) = \begin{cases} q(j|i,a)/\Lambda_N, & j \neq i \\ [\Lambda_N + q(i|i,a)]/\Lambda_N, & j = i . \end{cases}$$

(Note that $q_N \geq 0$ iff $\Lambda_N \geq \Lambda$. Naturally, we are only interested in $\mathcal{P}_N$ with $\Lambda_N \geq \Lambda$, and although we will label problems $\mathcal{P}_1, \mathcal{P}_2, \ldots$, it is understood that the first few $\mathcal{P}_N$ are, when necessary, to be omitted.)

If π is piecewise constant, then we can define the action selected by policy π in $\mathcal{P}_N$ when the current state is i and n transitions remain to be $\pi(n/\Lambda_N, i)$, and the return of policy π in $\mathcal{P}_N$ from state i when n transitions remain is denoted by $V_{\alpha,n,N}(\pi,i)$. However, if π is not piecewise constant, this definition could lead to executing actions in $\mathcal{P}_N$ which only rarely are executed in $\mathcal{P}$. For example, if $T = 1$, $S = \{0\}$, $r(0,i) = i$, and $\pi(t) = 1$ if t is rational and 0 otherwise, then $V_{0,1}(\pi) = 0$, yet $V_{0,2^N,N}(\pi) \equiv 1$. In view of this, we employ the following definition. With probability $\pi_{n,N}(i,a)$, action a is selected by policy π in $\mathcal{P}_N$ when the current state is i and n transitions remain where $\pi_{n,N}(i,a)$ is the Lebesgue measure of $\{t{:}\pi(t,i) = a,\ t \,\epsilon\, [\frac{2^N-n}{\Lambda_N}, \frac{2^N-n+1}{\Lambda_N}]\}$. If π is piecewise constant, then $\pi_{n,N}(i,a)$ converges to 1 [0] if $\pi(n/\Lambda_N,i) = a$ [$\neq a$] uniformly in i, a, and n as $N \to \infty$, so that the two definitions lead to the same asymptotic behavior (in n and N) of $V_{\alpha,n,N}(\pi)$. Similarly, we define the return function $V_{\alpha,n,N}$ for $\mathcal{P}_N$ by

$$V_{\alpha,n,N}(i) = \sup_{\pi} V_{\alpha,n,N}(\pi,i), \quad n = 0,1,2,\ldots,2^N .$$

An alternative and more useful formula for $V_{\alpha,n,N}$ is ($V_{\alpha,0,N} \equiv 0$)

$$V_{\alpha,n,N}(i) = \frac{1}{\alpha+\Lambda_N} \max_{a \epsilon A_i} \{r(i,a) + \sum_{j=0}^{\infty} q(j|i,a)V_{\alpha,n-1,N}(j)\} + \frac{\Lambda_N}{\alpha+\Lambda_N} V_{\alpha,n-1,N}(i) .$$

Finally, given $0 \leq t \leq T$, let $\langle t_N \rangle$ be a nondecreasing sequence with $t_N \leq 2^N$ and $t_N/\Lambda_N \to t$.

For the case $T = \infty$, P_N is defined as above except that we set $\Lambda_N = 2^N\Lambda$, take P_N to have infinitely many periods rather than 2^N, and let $\pi_{n,N}(i,\cdot)$ be the action prescribed by π in P_N for the n^{th} transition rather than when n transitions remain. In this case, we write $V_{\alpha,N}(\pi)$ and $V_{\alpha,N}$ for the return of π in P_N and the return function itself.

III. FINITE HORIZON RESULTS

We begin by demonstrating that $V_{\alpha,t_N,N}(\pi)$ converges to $V_{\alpha,t}(\pi)$ for every π and that $V_{\alpha,t_N,N}$ converges to $V_{\alpha,t}$. Next, we make use of $<\mathcal{P}_N>$ to construct a policy π^* defined on the diadic rationals in $[0,T]$ with the property that $V_{\alpha,t_N,N}(\pi^*) - V_{\alpha,t_N,N}$ converges (on some subsequence) to 0. Roughly speaking, π^* is optimal on a countable dense set of $[0,T]$. By imposing an additional structural condition on $<\mathcal{P}_N>$ -- and hence on $\mathcal{P}$ -- π^* has an (essentially) unique extension and is optimal for $\mathcal{P}$. In addition, we show that the optimal return function uniquely satisfies the appropriate functional equation.

LEMMA 1: If π is piecewise constant, then

$$V_{\alpha,t_N,N}(\pi) \to V_{\alpha,t}(\pi) .$$

Proof: Throughout the proof, the initial state i is fixed, and we employ the special definition for piecewise constant policies in $\mathcal{P}_N$. Let $\varepsilon > 0$ be given. To begin, define B_J to be the event that during $[0,T]$ the set $\{0,1,\ldots,i+J\}$ of states is left by the process induced by π, either in $\mathcal{P}$ or in some $\mathcal{P}_N$. Now Assumptions 1-3 ensure (see [8]) that $C < \infty$, where

$$C = \max\ \{1;\ V_{\alpha,t}(\pi,i);\ \sup_N V_{\alpha,T_N,N}(\pi,i)\}\ ,$$

so we can fix J so that $P(B_J)C < \varepsilon$.

Let $S_N(\omega) \subseteq [t,\infty)$ [1] be the set of times during which the actions specified by π and those induced by π in $\mathcal{P}_N$ are different. We claim

[1] While the expected duration of problem N is T, the realized duration may, in fact, be much larger than T.

that for any $\gamma > 0$

$$(5) \qquad P(L_N < \gamma | \tilde{B}_J) \to 1 \quad \text{as} \quad N \to \infty ,$$

where $L_N(\omega)$ is simply the Lebesgue measure of $S_N(\omega)$. (As $S_N(\omega)$ is, with probability 1, merely a finite union of intervals, L_N exists and is itself P-measurable.)

To begin, let $t = t_0 < t_1 < \ldots < t_M \equiv T$ be an enumeration of the set of times when the action specified by π from any state $j \leq i+J$ changes. Define $k(N,j)$ to be the smallest nonnegative integer such that $t + k(N,j)/\Lambda_N \geq t_j$. Take N sufficiently large so that

$$D_N \equiv \sum_{j=0}^{M} (t + k(N,j)/\Lambda_N - t_j) < \gamma/2 .$$

(The quantity D_N represents an upper bound on L_N if each period in P_N had length precisely $1/\Lambda_N$, its mean.)

Let $t_{N,j}$ be the sum of $k(N,j) - k(N,j-1)$ independent exponential random variables each having parameter Λ_N and assume that the components of $\{t_{N,j}\}$ are also independent. Next, define $T_{N,j} = t_{N,1} + \ldots + t_{N,j}$. The claim is established upon verifying that

$$P(\sum_{j=1}^{M} |T_{N,j} - E(T_{N,j})| < \gamma/2) \to 1 \quad \text{as} \quad N \to \infty ,$$

which follows, using the Chebychev Inequality, by noting that

$$P(|t_{N,j} - E(t_{N,j})| < \frac{\gamma}{2M^2}) \geq 1 - \frac{4M^4}{\gamma^2} \operatorname{Var}(t_{N,j})$$

$$= 1 - \frac{4M^4}{\gamma^2} \cdot \frac{k(N,j) - k(N,j-1)}{\Lambda_N^2} = 1 - \frac{4M^4}{\gamma^2} \cdot \frac{k(N,j) - k(N,j-1)}{N} \cdot \frac{T^2}{N}$$

$$\geq 1 - \frac{4M^4T^2}{\gamma^2} \cdot \frac{1}{N} .$$

Let S_N be that subset of the sample space such that for all s in $[t,T]$ we have $X_s = X_{N,s}$, where X_s and $X_{N,s}$ are the states of the process at time s in P and P_N, respectively.

Since the maximum rate at which the state can change is $\Lambda < \infty$, $P(S_N | L_N < \beta; \tilde{B}_J) \geq e^{-\Lambda\beta}$. This fact coupled with (5) enables us to conclude that

$$(6) \qquad \limsup_{N\to\infty} P(\tilde{B}_J \cap \widetilde{(S_N \cap L_N < \gamma)}) \leq P(\tilde{B}_J)(1 - e^{-\Lambda\gamma}) \ .$$

The absolute difference in cost in P and P_N for any ω in $L_N < \gamma \cap S_N \cap \tilde{B}_J$ is at most $\gamma \max\{|r(n,a)| : n \leq i+J,\ a \in A_n\} \equiv \gamma R$. Similarly, this difference for $\omega \in \tilde{B}_J \cap \widetilde{(S_N \cap L_N < \gamma)}$ is at most RT. Now choose γ so that $\gamma R < \varepsilon$ and $RT(1-e^{-\Lambda\gamma}) < \varepsilon$. Consequently, for N sufficiently large,

$$|V_{\alpha,t_N,N}(\pi,i) - V_{\alpha,t}(\pi,i)| < P(B_J)C + \gamma R + RT(1 - e^{-\Lambda\gamma}) < 3\varepsilon \ .$$

Q.E.D.

THEOREM 2. For any policy π,

$$V_{\alpha,t_N,N}(\pi) \to V_{\alpha,t}(\pi) \ .$$

Proof: Since π is measurable, there is a piecewise constant policy $\hat{\pi}$ with the property that for each i the measure of the subset ν_i of $[0,T]$ such that $\hat{\pi}(i,t) \neq \pi(i,t)$ is less than $\varepsilon/2^i$. Noting that

$$|V_{\alpha,t_N,N}(\pi,i) - V_{\alpha,t}(\pi,i)| \leq |V_{\alpha,t_N,N}(\pi,i) - V_{\alpha,t_N,N}(\hat{\pi},i)|$$

$$+ |V_{\alpha,t_N,N}(\hat{\pi},i) - V_{\alpha,t}(\hat{\pi},i)| + |V_{\alpha,t}(\hat{\pi},i) - V_{\alpha,t}(\pi,i)| \ ,$$

we need only verify that each of these three terms is small for N large.

Lemma 1 states that the middle term converges to zero. The third term is small because the probability that different sample paths will result is at most $1-e^{-\Lambda 2\varepsilon} \approx 2\Lambda\varepsilon$, so that the arguments of the proof of Lemma 1 suffice. To see that the first term is small, merely observe that $1-e^{-2\Lambda\varepsilon}$ is, for N sufficiently large, an upper bound on the probability that the sample paths are not exactly the same. When the sample paths are the same, the actions selected are the same except perhaps for an effective time interval of measure 4ε. Thus, the arguments of the proof of Lemma 1 again suffice.

Q.E.D.

LEMMA 3. For each t, $V_{\alpha,t_N,N} \to V_{\alpha,t}$.

Proof: From Theorem 2, we can choose π with $V_{\alpha,t}(\pi) > V_{\alpha,t} - \varepsilon$ so that

$$V_{\alpha,t_N,N}(i) \geq V_{\alpha,t_N,N}(\pi,i) \to V_{\alpha,t}(\pi,i) .$$

Thus, $\liminf_{N\to\infty} V_{\alpha,t_N,N} \geq V_{\alpha,t}$.

To see that $\limsup_{N\to\infty} V_{\alpha,t_N,N} \leq V_{\alpha,t}$, observe that $V_{\alpha,t_N,N} = V_{\alpha,t_N,N}(\pi_N)$ and we can, without loss of generality, take π_N to be constant on the 2^N intervals $(\frac{j}{\Lambda_N}, \frac{j+1}{\Lambda_N})$, $j = 0,1,\ldots,2^N-1$. Since $V_{\alpha,t_N,N}(\pi_N) = V_{\alpha,t_N/\Lambda_N}(\pi_N) \leq V_{\alpha,t_N/\Lambda_N}$ and $t_N/\Lambda_N \to t$, the continuity of $V_{\alpha,s}$ in s yields

$$\limsup_{N\to\infty} V_{\alpha,t_N,N} \leq \limsup_{N\to\infty} V_{\alpha,t_N/N} = V_{\alpha,t} .$$

Q.E.D.

We now show how to construct a policy π^* that is optimal for problem $\mathcal{P}$. To begin, for each i and each point in the set D of diadic rationals in $[0,T]$, we define π^* so that actions are selected in accord with an optimal policy for infinitely many $\mathcal{P}_N$. Assuming the presence of a certain structural property given below, we extend the definition of π^* from the dense set D to all of $[0,T]$ and then verify that π^* is optimal for $\mathcal{P}$.

Define $D_N = \{\frac{j}{2^N}T : j = 0,1,\ldots,2^N\}$, note that $D_N \nearrow D$, and for each $t \in D$, say $t = jT/2^M$, let $A_{i,t}$ be that set of actions which, for infinitely many N, is optimal from state i in $\mathcal{P}_N$ when $j2^{N-M}$ transitions remain. By diagonalization and the fact that each A_i is finite, we can find a subsequence $\langle\eta_N\rangle$ with the following property. For each $i \le N$ and each $t \in D_N$ (say $t = jT/2^M$), there is a subset $\hat{A}_{i,t}$ of $A_{i,t}$ such that $\hat{a} \in \hat{A}_{i,t}$ means that $\hat{a}$ is optimal in $\mathcal{P}_{\eta_k}$ when $j2^{\eta_k - M}$ periods remain for all $k \ge N$; $N = 1,2,\ldots$.

We say that $\langle\mathcal{P}_N\rangle$ is <u>connected</u> if for each state i and each N the optimality of action d when n and m periods remain with $n < m$ implies that d is also optimal when $n+1, n+2, \ldots,$ and $m-1$ periods remain.

Order the elements of each A_i and set $\pi^*(i,t)$ to be the minimal element of $\hat{A}_{i,t}$ for $t \in D$ and $i \in S$. Assuming that $\langle\mathcal{P}_N\rangle$ is connected, we can define the policy π^* to be the unique left-continuous extension. It is worth noting that π^* inherits all structural properties -- such as using faster rates from higher states -- present in $\langle\mathcal{P}_N\rangle$, including the property of connectedness. We will make extensive use of this in the applications.

THEOREM 4. If $<P_N>$ is connected, then the policy π^* is optimal for P. In particular, π^* is peicewise constant and for each i has at most card A_i switches.

Proof: Let i be the initial state. Since $<P_N>$ and π^* are connected, the relative frequency of periods in P_{n_N} when π^* is not using an optimal action from states $\{0,1,\ldots,i+J\}$ is, for each sample path, at most $f_N = \sum_{k=0}^{i+J} \text{card } A_k/2^N$, where $J < N$ is chosen so that the probability of reaching a state larger than $i+J$ when using either the optimal policy or π^* in P_{n_N} is less than ε for all N. As $f_N \to 0$, the idea of the proof of Lemma 1 suffices to yield $V_{\alpha,t_{n_N},n_N}(\pi^*,i) - V_{\alpha,t_{n_N},n_N}(i) \to 0$, which, coupled with Theorem 2 and Lemma 3, establishes the optimality of π^*.

Q.E.D.

REMARK. If instead of $<P_N>$ connected we assume: for each i

$$B_i \equiv \sup_N \{\# \text{ of periods in which the optimal action differs from preceding period's action for } P_N \text{ and state } i\} < \infty ,$$

then the proof works with B_i replacing card A_k and π^* is piecewise constant but not necessarily connected.

Restricting attention to the case $\alpha = 0$ and S finite, Miller [11] provided the following necessary and sufficient condition for optimality. The method of proof used here is Miller's.

THEOREM 5. A necessary and sufficient condition for a policy π to be optimal is that for almost all $t \in [0,T]$,

(7) $\quad r(f) + Q(f)\,\psi(t)$

is maximized over the set F of maps from S to A by $\pi(t)$, where $q(j|i,f(i))$ is the ij^{th} element of $Q(f)$ and the column vector $\psi(t)$ is the unique absolutely continuous solution to

$$(8) \qquad -\dot{\psi}(t) = r(\pi(t)) + Q(\pi(t))\psi(t) - \alpha\psi(t); \quad \psi(T) \equiv 0 .$$

Moreover, the solution ψ to Equation 8 satisfies

$$(9) \qquad \psi(t) = \int_t^T e^{-\alpha(s-t)} P(t,s)r(\pi(s))ds = V_{\alpha,t}(\pi) .$$

Proof: The proof is nearly that of Miller [11], as the case $\alpha > 0$ is straightforward (see [9] for details). To extend the proof to the case S countable, it sufficies to demonstrate the uniqueness and existence of a solution to Equation 8. To begin, write $r(s)$ and $Q(s)$ for $r(\pi(s))$ and $Q(\pi(s))$, define the Banach space B by

$$B \equiv \{\langle u(i)\rangle_{i\varepsilon S} : \sup |u(i)|/(i\vee 1)^m < \infty\} ,$$

and let M be the (metric) space of continuous functions on $[0,T]$ into B with metric $d(x,y) = \max_{0\leq s\leq T} \|x(s)-y(s)\|$, where $\|\;\|$ is the norm in B. Next, define the map $F:M\to M$ by

$$[Fx(t)]_i = z_i(T) + \int_t^T \{r_i(s) + (Q(s)x(s))_i - \alpha x_i(s)\}ds .$$

It will suffice to show that F is well defined and has a unique fixed point. Noting that

$$(Q(s)x(s))_i = \sum_{j=0}^{\infty} q(j|i,\pi(s,i))x_j(s) = \lim_{n\to\infty} \sum_{j=0}^{n} q(j|i,\pi(s,i))x_j(s) ,$$

we see that $(Q(s)x(s))_i$ is measurable since $x_j(s)$ is continuous, $q(j|i,\pi(s,i))$ is measurable, and the limit of measurable functions is measurable (the sum converges by Assumption 3). Hence, F is well defined.

From Assumption 3, it follows that there is a uniform bound on $||Q(f)||$ over F, say D, so that $||Q(s)|| \leq D$, $0 \leq s \leq T$. Consequently, $d(F(x),F(y)) \leq TD$, so that F is a contraction and has a unique fixed point if $TD < 1$. If TD is not less than one, then merely choose a time T' for which $T'D < 1$ and piece together the desired solution by appropriate choice of the initial value $z_i(T)$.

Q.E.D.

In view of Equation 9 and Theorem 5, we are lead to inquire whether the optimal return function V_α satisfies

$$(10) \qquad -\dot{\psi}(t) = \max_{f \in F} \{r(f) + Q(f)\psi(t)\} - \alpha\psi(t), \qquad \psi(T) \equiv 0 .$$

<u>COROLLARY 6</u>. There is an optimal policy, and the optimal return function is the unique solution (in M) to (10).

<u>Proof</u>: Let B and M be given as in the proof of Theorem 5 and define $\hat{F}:M \to M$ by

$$(11) \qquad [\hat{F}x(t)]_i = \int_t^T \max_{a \in A_i} \{r(i,a) + \sum_{j=0}^{\infty} q(j|i,a)x_j(s) - \alpha x_i(s)\}ds .$$

Fix $x \in M$, i, and $a \in A_i$. We claim that the maximand is continuous. To see this, choose $t \in [0,T]$, let $\varepsilon > 0$ be given and choose $\delta > 0$ so that $||x(s)-x(t)|| < \varepsilon/(i+b)^m$ whenever $|s-t| < \delta$. Then by Assumption 3 we have

$$\begin{aligned}&\left|\Sigma q(j|i,a)x_j(s) - \Sigma q(j|i,a)x_j(t)\right| \\ &\qquad \leq \Sigma q(j|i,a)\left|x_j(s) - x_j(t)\right| < \Sigma q(j|i,a)\ \varepsilon\ (j\vee 1)^m/(i+b)^m \leq \varepsilon\ ,\end{aligned}$$

justifying our claim. Consequently, for each i, the action attaining the maximum can be chosen measurably so that $\hat{F}$ is well defined. Finally, $\hat{F}$ has a unique fixed point which, by Theorem 5, is V_α.

Q.E.D.

IV. INFINITE HORIZON RESULTS

The infinite horizon problem is considerably more straightforward than the case $T < \infty$. In fact, the existence of an optimal stationary policy follows from three simple observations. First, for any stationary policy $f \in F$ the return of f is the same in P and $<P_N>$; that is

$$(12) \qquad V_\alpha(f) = V_{\alpha,N}(f), \quad \text{all} \quad N .$$

Second, by known results concerning semi-Markov decision processes (see Theorem 1 of [8]), there is an $f_\alpha \in F$ such that

$$(13) \qquad V_{\alpha,N}(f_\alpha) = \sup V_{\alpha,N}(\pi), \quad \text{all} \quad N ,$$

where the sup is taken over all policies, including randomized and history dependent policies. Third, for each $\varepsilon > 0$ and initial state i, we can (in view of Assumptions 1, 2, 3) find a time $T_{\varepsilon,i} < \infty$ such that the total expected α-discounted reward received after time $T_{\varepsilon,i}$ when starting from state i is less than ε. Coupling this with Theorem 2 yields

$$(14) \qquad V_{\alpha,N}(\pi) \to V_\alpha(\pi), \quad \text{for each policy} \quad \pi .$$

THEOREM 7. For each $\alpha > 0$, there is an $f_\alpha \in F$ such that $V_\alpha(f_\alpha) = V_\alpha$.

Let $V(\pi,T,i)$ denote the total expected reward earned by time T when employing policy π and starting from state i, and define $\overline{V}(\pi)$, the average expected return per unit time of policy π, by

$$\overline{V}(\pi) = \liminf_{T\to\infty} V(\pi,T)/T .$$

Then if there is an $f^* \in F$ and a sequence $\alpha_n \searrow 0$ with $V_{\alpha_n}(f^*) = V_{\alpha_n}$, we have, employing a standard Abelian result (see Lemma 1 of [6]),

$$\overline{V}(\pi) = \liminf_{T\to\infty} V(\pi,T) \leq \liminf_{\alpha\to 0^+} \alpha V_\alpha(\pi) \leq \liminf_{n\to\infty} \alpha_n V_{\alpha_n}(\pi)$$

$$\leq \liminf_{n\to\infty} \alpha_n V_{\alpha_n}(f^*) = \overline{V}(f^*) \ .$$

The existence of such an f* is ensured if S is finite. More important, however, is the fact that the existence of such an f* can often be verified in countable state problems arising in the context of specific applications (see Example 2).

V. APPLICATIONS

The following three examples taken from the queueing optimization literature illustrate the use of our approach.

EXAMPLE 1. Optimal Customer Selection in an M/M/c Queue (Miller-Cramer-Lippman)

We consider the problem of determining which customers to admit into the system so as to maximize the expected α-discounted reward earned over a finite horizon of length T in an M/M/c queue with finite queue capacity Q. Each customer class, $1 \leq k \leq K < \infty$, is distinguished only by the reward r_k associated with acceptance of a class k customer into the queue and the Poisson arrival rate λ_k. Each of the c exponential servers has rate μ. For convenience, label the customer classes so that $0 < r_1 < r_2 < \ldots < r_K$.

Even though the rewards are received in lump sums and not as rates, the following clever problem representation due to Miller [13] permits formulation as a continuous time Markov decision process. Take $S = \{0,1,2,\ldots,c+Q\}$ so that being in state i means that there are i customers in the system. For $i < c+Q$, take A_i to be the power set of $\{1,2,\ldots,K\}$ so that each action $a \in A_i$, a subset of $\{1,2,\ldots,K\}$, merely stipulates which customer classes are to be admitted ($A_{c+Q} = \emptyset$). Then the reward and transition rates are given by $r(a,i) = \sum_{j\varepsilon a} \lambda_j r_j$, $q(i+1|i,a) = \sum_{j\varepsilon a} \lambda_j$, $q(i-1|i,a) = (i\wedge c)\mu$, and $q(j|i,a) = 0$, $j \neq i-1,i,i+1$.

By considering the appropriate semi-Markov version of this problem (with any rate $\Lambda \geq c\mu + \sum_1^K \lambda_j$), it can be shown (see Theorems 4, 5, 6 of [7]) that the minimal reward accepted when the discount factor is α, there are i customers in the system, and n transitions remain is an increasing function of $1/\alpha$, i, and n. Consequently, the approximating problems are connected so we can conclude from Theorem 4 that associated with each $\alpha \geq 0$ there is an optimal policy π_α with the following property: the set $\pi_\alpha(i,t)$ of customer classes accepted from state i at time t is nonincreasing in T-t, i, and $1/\alpha$; moreover, $\pi_\alpha(i,t)$ is of the form $\{j,j+1,\ldots,K\}$. (This result was obtained by Miller [Theorem 7.3, 10] for the case $Q = \alpha = 0$.)

EXAMPLE 2. An M/M/1 Queue with Variable Service Rate (Crabill-Sabeti)

The problem posed by Crabill [2] was to determine the service rate $\mu \in \{\mu_j\}_1^K$ to employ so as to minimize the expected average cost per unit time in an M/M/1 infinite capacity queue with arrival rate λ in the presence of a holding cost h per customer per unit time and a service cost rate c_j associated with the service rate μ_j.[2/] For convenience, label $0 < \mu_1 < \mu_2 < \ldots < \mu_K$ and assume, as is reasonable, that $c_1 < c_2 < \ldots < c_K$. Our interest is in minimizing the expected α-discounted cost incurred during a horizon of length $T \leq +\infty$.

Here, the natural formulation -- the state is the number of customers in the system and the action is the rate μ_j to employ -- suffices. Utilizing results from [7], we can, as in Example 1, conclude that for $T < \infty$ and each $\alpha \geq 0$ there is an optimal policy π_α such that $\pi_\alpha(i,t)$, the optimal rate from state i when t time units remain, is nondecreasing in i, t, and $1/\alpha$. Furthermore, Theorem 7 enables us to assert that the optimal rate to employ when $T = \infty$ is a nondecreasing function of both

[2/] A reward for service completions can also be included (see [7, p. 36]).

$1/\alpha$ and i. (Note that with the natural formulation in the infinite horizon version of Example 1 there are no actions available unless an arrival occurs, whereas actions are continuously available in this variable service rate model.)

EXAMPLE 3. Optimal Admission to an M/M/1 Queue (Emmons)

We consider the single server version of Emmons' [3] M/M/c infinite capacity queue with arrival rate λ and service rate μ. The problem is to dynamically determine whether or not to admit arriving customers into the system, counterbalancing the reward r received for admitting a customer against the possible overtime service cost which is incurred beginning at time T and continuing until the system is empty. Customers cannot be admitted after time T. Our goal is to maximize the expected α-discounted net profit.

In order to represent the problem as a continuous time Markov decision process, Emmons introduced a non-zero terminal condition to incorporate the overtime cost and utilized Miller's method to handle the rewards associated with customer acceptance. Instead of assuming that the overtime service cost $c(\tau)$ associated with closing τ units late is linear, we assume that $S(i)$, the expected α-discounted value of overtime service costs associated with i customers present at time T, is convex in i. For example, $c(\tau) = \tau e^{\beta\tau}$ with $\beta \geq \alpha$ yields the desired convexity.

In order to verify that the approximating problems are connected, we need the following two results. The first, Lemma 8, states that there is a sequence $\langle i_n \rangle$ of critical numbers with the property that it is optimal to accept a customer when n transitions remain and i customers are currently in the system iff $i < i_n$. The second result, Lemma 9, states that $i_0 \leq i_1 \leq i_2 \leq \ldots$. In view of Lemmas 8 and 9, we can employ Theorem 4 as in Examples 1 and 2 to verify that there is an optimal policy for P characterized by a sequence $T \geq t_0 \geq t_1 \geq \ldots \geq 0$ of critical numbers as follows: a customer seeking admission at time t when there are i customers in the system will be admitted iff $t < t_i$. (It is likely that $t_k = 0$ for some $k < \infty$ and $t_0 = T$ iff $r < S(1)$.)

Fix $\Lambda \geq \lambda+\mu$ and $\alpha \geq 0$. Then $V_n(i)$, the n-period return starting from state i satisfies ($V_0(i) = -S(i)$)

$$V_{n+1}(i) = \frac{1}{\alpha+\Lambda} \max \{\lambda r + \lambda V_n(i+1) + \mu V_n(i-1) + (\Lambda-\lambda-\mu)V_n(i) ;\\ \mu V_n(i-1) + (\Lambda-\mu)V_n(i)\} ,$$

or

$$V_{n+1}(i) = \frac{1}{\alpha+\Lambda} \{\lambda \max [r + v_n(i+1);0] + \mu V_n(i-1) + (\Lambda-\mu)V_n(i)\} ,$$

where $v_n(i+1) = V_n(i+1) - V_n(i)$.

<u>LEMMA 8</u>. For each n, $V_n(\cdot)$ is concave; that is, $v_n(i) \geq v_n(i+1)$ all i.

<u>LEMMA 9</u>. For each $n \geq 0$ and each $i \geq 1$, $v_{n+1}(i) \geq v_n(i)$.

The proofs of Lemmas 8 and 9 are given in [9].

REFERENCES

1. Cramer, M., "Optimal Customer Selection in Exponential Queues," ORC 71-24, Operations Research Center, University of California, Berkeley, 1971.

2. Crabill, T.B., "Optimal Control of a Queue with Variable Service Rates," Ph.D. Dissertation, Cornell University, 1968.

3. Emmons, H., "The Optimal Admission Policy to a Multiserver Queue with Finite Horizon," J. Appl. Prob., 9, 103-116 (1972).

4. Harrison, J.M., "Discrete Dynamic Programming with Unbounded Rewards," Ann. Math. Stat., 43, 636-644 (1972).

5. Kakumanu, P., "Continuously Discounted Markov Decision Model with Countable State and Action Space," Ann. Math. Stat., 42, 919-926 (1971).

6. Lippman, S.A., "Semi-Markov Decision Processes with Unbounded Rewards," Management Science, 19, 717-731 (1973).

7. Lippman, S.A., "A New Technique in the Optimization of Exponential Queueing Systems," Working Paper No. 211, Western Management Science Institute, UCLA, October 1973.

8. Lippman, S.A., "On Dynamic Programming with Unbounded Rewards," Working Paper No. 212, Western Management Science Institute, UCLA, November 1973.

9. Lippman, S.A., "Countable State Continuous Time Dynamic Programming with Structure," Discussion Paper No. 42, Operations Research Study Center, Graduate School of Management, UCLA, December 1973.

10. Miller, B.L., "Finite State Continuous Time Markov Decision Processes with Applications to a Class of Optimization Problems in Queueing Theory," Technical Report No. 15, Department of Operations Research, Stanford University, March 10, 1967.

11. Miller, B.L., "Finite State Continuous Time Markov Decision Processes with a Finite Planning Horizon," *SIAM J. on Control*, *6*, 266-280 (1968).

12. Miller, B.L., "Finite State Continuous Time Markov Decision Processes with an Infinite Planning Horizon," *J. Math. Anal. and Appl.*, *22*, 552-269 (1968).

13. Miller, B.L., "A Queueing Reward System with Several Customer Classes," *Management Science*, *16*, 234-245 (1969).

14. Prabu, N. and S. Stidham, Jr., "Optimal Control of Queueing Systems," Technical Report No. 186, Department of Operations Research, Cornell University, 1973.

15. Sabeti, H., "Optinal Decision in Queueing," Technical Report No. 12, Operations Research Center, University of California, Berkeley, April 1970.

16. Veinott, A.F., Jr., "Discrete Dynamic Programming with Sensitive Discount Optimality Criteria," *Ann. Math. Stat.*, *40*, 1635-1660 (1969).